Allergen Chip Challenge Writeup

Rakesh Jarupula
21/07/2023

Intro:

Hello, I am Rakesh Jarupula. I have graduated in Electrical Engineering from National Institute of Technology Silchar (IN). I was a Data Scientist at BetterPlace. I acquired my skills mainly through Coursera and Kaggle competitions. Also, I am a 2X Kaggle expert.

Interest:

When I saw the competition description for the first time, I felt like the competition is a challenging opportunity to apply my data science skills and knowledge in a practical setting. It allowed me to explore various machine learning models, feature engineering techniques, and data analysis strategies to improve model performance. The Allergen Chip challenge provided a unique opportunity to work on a task that has direct implications for public health and well-being. It provided an opportunity to enhance my problem-solving abilities, work under time constraints. The challenge pushed you to think critically, analyze data thoroughly, and implement sophisticated models to achieve my best possible performance.

Approach:

The biggest boost to my score is feature engineering. I tried most of my time to understand the data and derive best possible features to handle the problem at hand. I will explain my approach and modeling decisions here:

Pre-Processing:

As per my observation of the data:

- Train and Test have different distribution.
- To reduce the difference, I dropped the rows that have 9 in ANY of the target columns.
- Dropped 'Food_Type_0' as it only has two non-missing instances.
- Replaced extreme values with 2nd highest value.
- Created an Excel sheet that maps different allergen proteins with their source and way of entering the body.

- Dropped some allergen proteins that doesn't help in predicting different allergies.
- Created new columns that correspond to the treatments that a patient has taken.

Feature Engineering:

I created new features based on both rows and columns.

Row based:

1. Zero counts
2. Missing value counts
3. Mean, Median, sum, std, maximum.

Column based:

1. Calculated sum, mean, min, median, std, maximum values in the columns that corresponds to Similar allergen.
2. Dropped old allergen proteins values and used only the derived column features.
3. Also treated the extreme values.

Model training:

I didn't use all the data for training models for predicting a particular target. Ex: For predicting the presence of Food allergy, I used the features that corresponds to the Food allergen proteins along with the meta features of the patients such as Age, Gender, Treatment types etc.

1. I trained different models for different targets instead of building a ChainClassifier. As the presence of One type of allergy may not have any effect on the other type of allergy as discussed in the forum. Also, tested training One with no use.
2. To deal with imbalance in the class labels I used scale_pos_weights parameters of the models.
3. Used Bayesian Search method find the Hyper-Parameters of the model. I used this technique since it uses cross-validation to find the best parameters, which is very important for unbiased and Robust model.
4. After finding the best parameters, I used Repeated Stratified KFold to train individual model in order to deal with class imbalance and to get robust prediction.
5. I used different threshold values for each target to deal with imbalance in the data.
6. Finally, I took the mean of the predictions of the 3 models in order to not overfit the data.

Resources:

In my opinion RAM and CPU are sufficient for this task. BUT I FACED CHALLEGES WITH RESPONSE OF THE SEVER DURING LAST FEW DAYS, WHICH HAD NARROWED MY APPROACHES TO THE PROBLEM. I could have even improved the score.